Обучение машинного перевода без параллельных текстов*

Иванов А. В. Бахтеев[1] ***О. Ю. Стрижов***[2] ***В. В.***
ivanov.aleksandr@phystech.edu
[1]Московский физико-технический институт
[2]Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН

Решается задача машинного перевода между двумя языками. В случае некоторых пар языков не удается построить тренировочную выборку из параллельных текстов достаточного размера. Для решения этой проблемы были предложены способы оптимизации моделей, основанные на автокодировщиках. Каждому предложению из обоих языков из параллельных текстов ставится в соотвествие некоторый "смысловой" вектор в скрытом пространстве. Модель оптимизируется для перевода с одного языка на другой при попытках восстановления изначальных предложений по их скрытому представлению. В данной работе проводится исследование предложенного подхода для перевода между двумя близкими языками: русским и украиским.

Ключевые слова: *перевод, автокодировщик.*

Введение

Выбор модели машинного обучения, используемой для перевода, зависит от спецефических особенностей пары языков. Так, использование глубоких нейронных сетей приводит к качественным результатам, но только в том случае, если количество параллельных предложений в обучающей выборке достаточно велико. В работах [1, 2] были достигнуты результаты для милионной выборки.

В случае, когда размер обучающей выборки не достаточен, для ее пополнения может быть использован результат предыдущих итераций обучаемой нейронной сети. Данный результат представлен в [3].

Во многих работах представлено решение задачи машинного перевода в том случае, когда количества параллельных предложений не достаточно для построения и оптимизации глубокой сети [4, 5, 6]. В данном подходе используются 2 типа моделей машинного обучения. Первый - рекуррентные нейронные сети LSTM [7] используются для перевода слов изначального языка в скрытое векторное пространство. Второй - дискриминатор, восстанавливающий исходное предложение по скрытому внутреннему представлению, являющемуся результатом работы первой сети.

Оптимизация проводится в состязательном режиме. Дискриминатор модели минимизирует разницу между скрытыми представлениями предложений из двух языков. Для борьбы с переобучением добавляется шум, который не дает возможности абсолютно точного восстановления текста по его представлению после обработки автокодировщиком. Шаг оптимизации состоит из двух частей. Изначально выбирается случайное предложение из исходного языка и кодируется с добавлением шума [8] после чего подаётся на вход дискриминатору. Далее аналогичная процедура со случайным предложением из второго языка. На второй стадии выбирается произвольное предложение из исходного языка, пе-

---

Научный руководитель: Стрижов В. В. Авторы: А.В. Грабовой, О.Ю. Бахтеев, В.В. Стрижов, Eric Gaussier, координатор Малиновский Г. С. Консультант: Бахтеев О. Ю.

реводится текущей моделью на конечный язык. На результат накладывается шум, и для зашумленного предложения выполняется полный обратный перевод, после чего вычисляется функция потерь. Затем аналогичные действия проводятся с произвольно выбранным предложением конечного языка.

Для перевода с русского на французский языки и обратно такой подход был продемонстрирован в [9].

В качестве эксперимента производится перевод предложений с русского языка на украинский. Качество результата оценивается с помощью метрики BLEU [10].

Постановка задачи

Рассматриваемая модель состоит из кодировщика $\mathfrak{U}$ и декодировщика $\mathfrak{d}$, и отвечающих соответственно за отображение предложений из обоих языков в латентное пространство и обратное отображение из латентного пространства в предложения первого или второго языка. Кодировщик и декодировщик реализованы в виде рекуррентных нейронных сетей. Будем рассматривать модели кодировщиков и декодировщиков для обоих языков: f^{src}, f^{tgt} и g^{src}, g^{tgt} соответственно. Модели имеют общие параметры, но отличаются входными словарями, разными для каждого языка. Пусть заданы следующие выборки: $\mathfrak{D}^{\text{src}} = [\mathbf{s}_1^{\text{src}}, \ldots, \mathbf{s}_{m_{\text{src}}}^{\text{src}}]$ - набор предложений из первого языка, $\mathfrak{D}^{\text{tgt}} = [\mathbf{s}_1^{\text{tgt}}, \ldots, \mathbf{s}_{m_{\text{tgt}}}^{\text{tgt}}]$ - набор предложений второго языка. Два предыдущих набора не обязаны быть параллельными. Так же есть валидационная выборка, являющаяся набором параллельных предложений. $\mathfrak{D}^{valid} = \{(\mathbf{s}_1^{\text{src}}, \mathbf{s}_1^{\text{tgt}}), \ldots, (\mathbf{s}_{m_{\text{valid}}}^{\text{src}}, \mathbf{s}_{m_{\text{valid}}}^{\text{tgt}})\}$.

Далее определеним функционалы, которые будут подвергнуты минимизации, которые и будут представлять собой функцию потерь для модели. Пусть σ - функция зашумления аргумента, применяемая перед началом обратного перевода модели. В качестве нормы может быть использована кросс-энтропия или стандартная L_2 норма..

$$L_{\text{sent}} = ||\mathbf{g}(\mathbf{f}(\sigma(x))) - x||$$

При пословном переводе функция потерь будет иметь вид

$$L_{\text{word}} = ||\mathbf{g}(\mathbf{f}(\mathbf{g}^{-1}(\mathbf{f}(x))) - x||$$

Оптимизация дискриминатора для того, чтобы он мог отличать представления различных языков в скрытом пространстве:

$$L_G = \log \mathbb{P}(\mathbf{f}_{\text{lang}} = 1|\mathbf{g}(x)) + \log \mathbb{P}(\mathbf{f}_{\text{lang}} = 2|\mathbf{g}(y))$$

Итоговая функция ошибки принимает следующий вид:

$$L = (L_{sent}, L_{word}, L_G)^T \cdot \mathbf{w} \to \min$$

$\mathbf{w}$ рассматривается как вектор весов "значимости"штрафа.

Литература

[1] Bilingual word embeddings for phrase-based machine translation / Will Y Zou, Richard Socher, Daniel Cer [и др.] // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013. С. 1393–1398.

[2] On the properties of neural machine translation: Encoder-decoder approaches / Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau [и др.] // arXiv preprint arXiv:1409.1259. 2014.

[3] Bertoldi Nicola, Federico Marcello. Domain adaptation for statistical machine translation with monolingual resources // Proceedings of the fourth workshop on statistical machine translation / Association for Computational Linguistics. 2009. С. 182–189.

[4] Google's neural machine translation system: Bridging the gap between human and machine translation / Yonghui Wu, Mike Schuster, Zhifeng Chen [и др.] // arXiv preprint arXiv:1609.08144. 2016.

[5] Sutskever Ilya, Vinyals Oriol, Le Quoc V. Sequence to sequence learning with neural networks // Advances in neural information processing systems. 2014. С. 3104–3112.

[6] Bahdanau Dzmitry, Cho Kyunghyun, Bengio Yoshua. Neural machine translation by jointly learning to align and translate // arXiv preprint arXiv:1409.0473. 2014.

[7] Graves Alex, Schmidhuber Jurgen. Framewise phoneme classification with bidirectional LSTM and other neural network architectures // Neural Networks. 2005. Т. 18, № 5-6. С. 602–610.

[8] Kim Yunsu, Ney Jiahui Geng Hermann. Improving Unsupervised Word-by-Word Translation with Language Model and Denoising Autoencoder. 2018.

[9] Lample Guillaume, Denoyer Ludovic, Ranzato Marc'Aurelio. Unsupervised machine translation using monolingual corpora only // arXiv preprint arXiv:1711.00043. 2017.

[10] BLEU: a method for automatic evaluation of machine translation / Kishore Papineni, Salim Roukos, Todd Ward [и др.] // Proceedings of the 40th annual meeting on association for computational linguistics / Association for Computational Linguistics. 2002. С. 311–318.

[11] Word translation without parallel data / Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato [и др.] // arXiv preprint arXiv:1710.04087. 2017.